Data Science

CODERHOUSE

Comisión 61790

Alumna: Méndez, Julieta Milagros

2025

Parte 1
Preprocesamiento de texto

¿Qué es el Procesamiento de Lenguaje Natural (NLP)?
Rama de la IA que se centra en la interacción entre las computadoras y el lenguaje humano.

¿Su objetivo?
Permitir que las máquinas comprendan, interpreten y generen lenguaje humano que sea natural para las personas (incluye texto escrito como lenguaje hablado)

Práctica: Análisis de reseñas

El dataset en estudio contiene más de mil críticas de películas y series españolas.

Ofrece una oportunidad para aplicar y experimentar técnicas de procesamiento de texto en nuestro idioma, promoviendo el desarrollo de proyectos, análisis y modelos específicos.

PASOS A SEGUIR:

1. PREPROCESAMIENTO
2. ANÁLISIS DE SENTIMIENTOS Y PALABRAS CLAVE
3. TRIGRAMAS MÁS FRECUENTES
4. CONCLUSIONES

1 PREPROCESAMIENTO

Pasos claves del preprocesamiento

TOKENIZACIÓN

Dividir el texto en partes más pequeñas, (palabras o frases: tokens). Ej.: "Muy buena película " → ["Muy", "buena", "película"]

STOPWORDS

Palabras muy comunes que no aportan mucho significado y se eliminan.
Ejemplo: "la", "de", "que", "y"

LEMATIZACIÓN

Reducir las palabras a su forma base o raíz.
Ejemplo: "cantando", "cantó", "cantaría" → "cantar"

1 PREPROCESAMIENTO

Conlusión del preprocesamiento

Reducción de ruido y generalización del texto: por lematización y remoción de stopwords, se logró reducir el texto a conceptos escenciales. Esto permite mejorar la calidad del análisis posterior (ej. análisis de sentimiento o clasificación).

Preservación de la opinión crítica: pese al filtro, los tokens finales siguen mostrando con claridad el tono negativo de las críticas. "Mediocre", "verguenza", "crítica", "falta", "culpa", "malo", muestran sentimientos desfavorables lo que confirmaría que el preprocesamiento no elimina lo relevante de las críticas.

Desbalance temático y tono en las reseñas: se observan bastantes palabras negativas y adjetivos con carga fuerte ("mediocre", "ofensiva", "doler", "verguenza") lo que indicaría que muchas reseñas puedan tener una polaridad negativa y esto podría influir en modelos de clasificación.

2

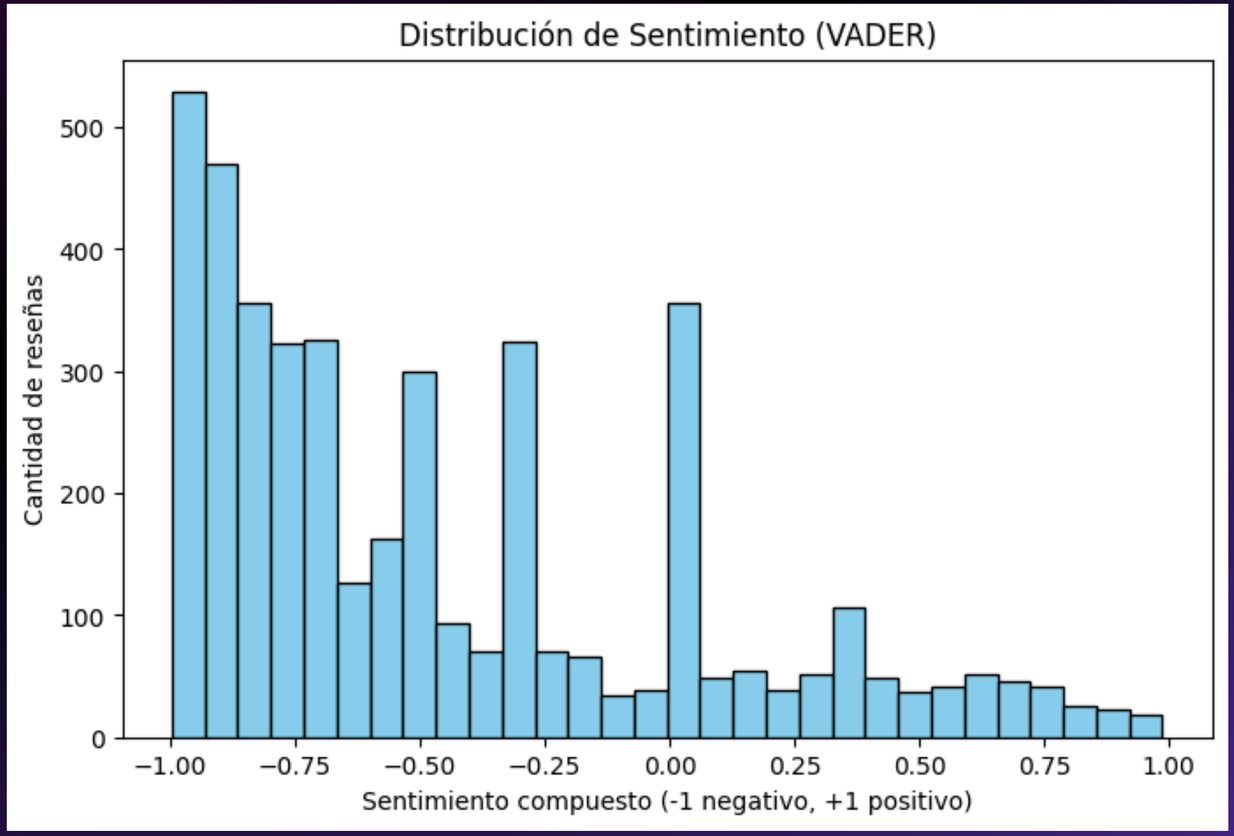
ANÁLISIS DE SENTIMIENTO Y PALABRAS CLAVE

ANÁLISIS DE SENTIMIENTO

Mide si un texto transmite una emoción negativa (-1), neutra (0) o positiva (1).

PALABRAS CLAVE

Identifica las palabras más importantes de un texto según su frecuencia y relevancia.



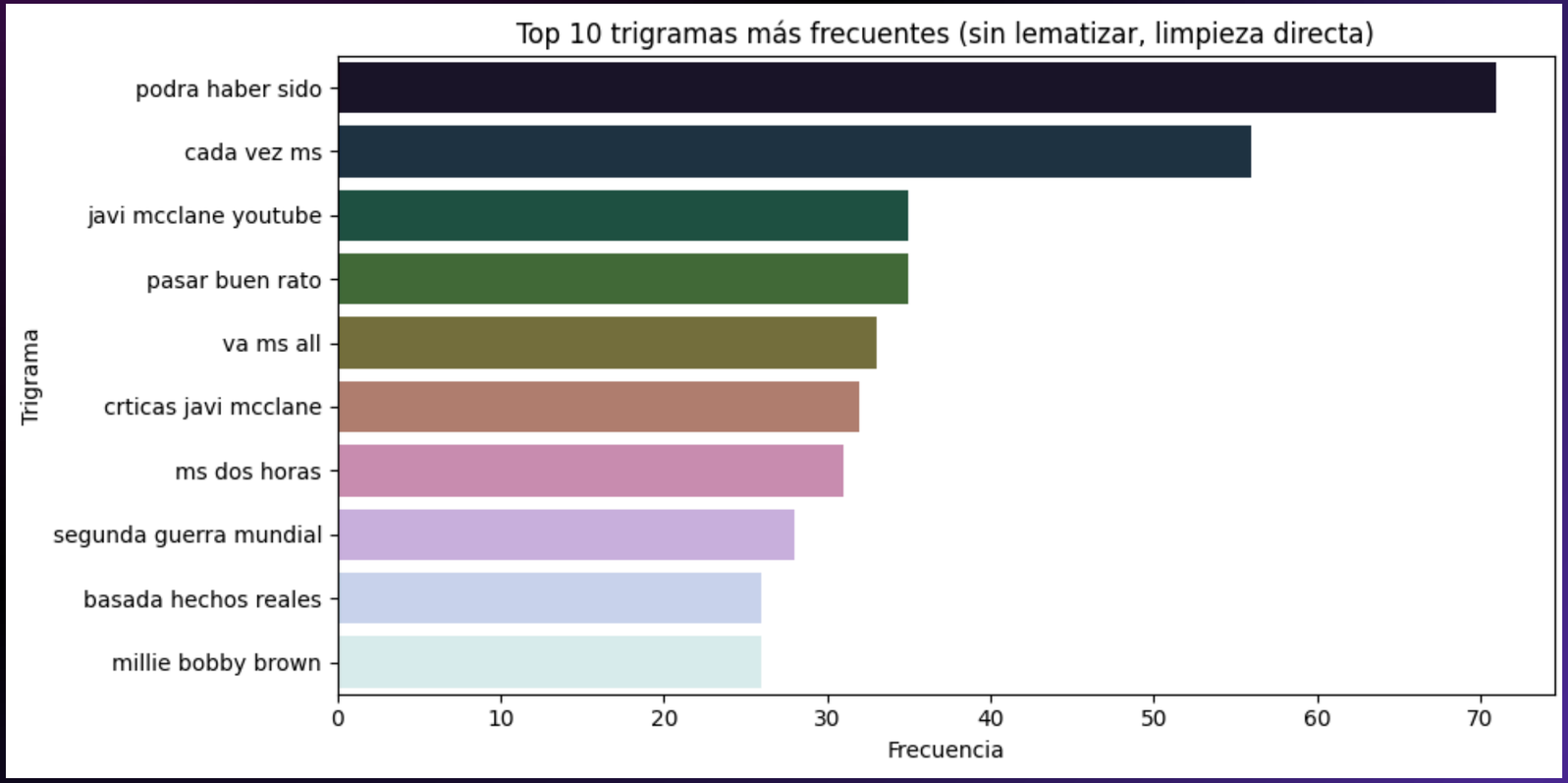
Varias críticas tienden a ser negativas. Pocas reseñas presentan valores positivos lo que puede sugerir que predominan opiniones negativas o que los ejemplos elegidos tienden a destacar lo negativo. Se observa desbalance en las clases.

Las palabras con mayor peso promedio son: "más", "serie", "película", "historia", "bien", "ver". Esto indica que el contenido de las críticas gira en torno a opiniones personales sobre las tramas, personajes y calidad general de las series/películas. Términos como "serie", "película", "historia", "personajes" sugiere que las críticas se centran en aspectos narrativos y de guion.

3

TRIGRAMAS MÁS FRECUENTES

Un trigrama es una secuencia de tres palabras que aparecen juntas en un texto.
Analizar los más frecuentes ayuda a detectar frases comunes o patrones de lenguaje.



"podrá haber sido", "cada vez más" o "pasar buen rato" -> expresiones comunes que aportan información emocional, típico de reseñas subjetivas.

"javi mcclane youtube" o "millie bobby brown" -> críticas dirigidas a actores/creadores específicos, útil si se quiere aplicar análisis de entidades o segmentar por tipo de contenido.

"segunda guerra mundial" o "basada hechos reales" -> son recurrentes dentro de las reseñas géneros o contextos históricos. Útil para clasificar temáticamente o identificar tendencias narrativas.
"va ms all" o "crticas javi mcclane" -> aún hay ruido en los datos. Una limpieza más profunda podría mejorar la calidad de los análisis posteriores.

Conclusiones sobre el dataset elegido

Se logró experimentar con herramientas de análisis textual y contribuir al desarrollo de recursos en español, idioma que aún presenta limitaciones en comparación con el inglés dentro del campo del NLP.

El proceso de preprocesamiento permitió limpiar el texto y reducirlo a sus componentes más relevantes, preservando el significado y facilitando los análisis posteriores. Como resultado se observó que luego los textos conservaban su tono crítico, reflejando opiniones fuertemente negativas, lo que también se confirmó con el análisis de sentimiento, donde la mayoría de los valores fueron negativos o cercanos a -1.

El cálculo de TF-IDF y el análisis de trigramas mostró que los términos más representativos están directamente relacionados con el contenido audiovisual, lo que confirma la coherencia temática del dataset.

Estos resultados muestran el valor de aplicar técnicas básicas de NLP para obtener información útil y para detectar desafíos específicos del idioma. Este trabajo podrá ser un primer paso hacia el desarrollo de modelos más robustos para el análisis de texto en español.

Parte 2
Deep learning

¿Qué es el Deep Learning?
Es un subconjunto del Machine Learning que se basa en el uso de redes neuronales artificiales para aprender de grandes cantidades de datos.

¿Su objetivo?
Lograr identificar patrones complejos y aprender representaciones de datos. Útiles para reconocimiento de imágenes, el procesamiento de lenguaje natural (NLP) y el análisis de datos no estructurados.

Práctica: Análisis de reseñas

Las opiniones de los usuarios juegan un papel clave en la percepción y mejora de productos y servicios. El dataset elegido contiene reseñas y valoraciones de usuarios sobre la aplicación de compras de Amazon, actualizadas de forma diaria.

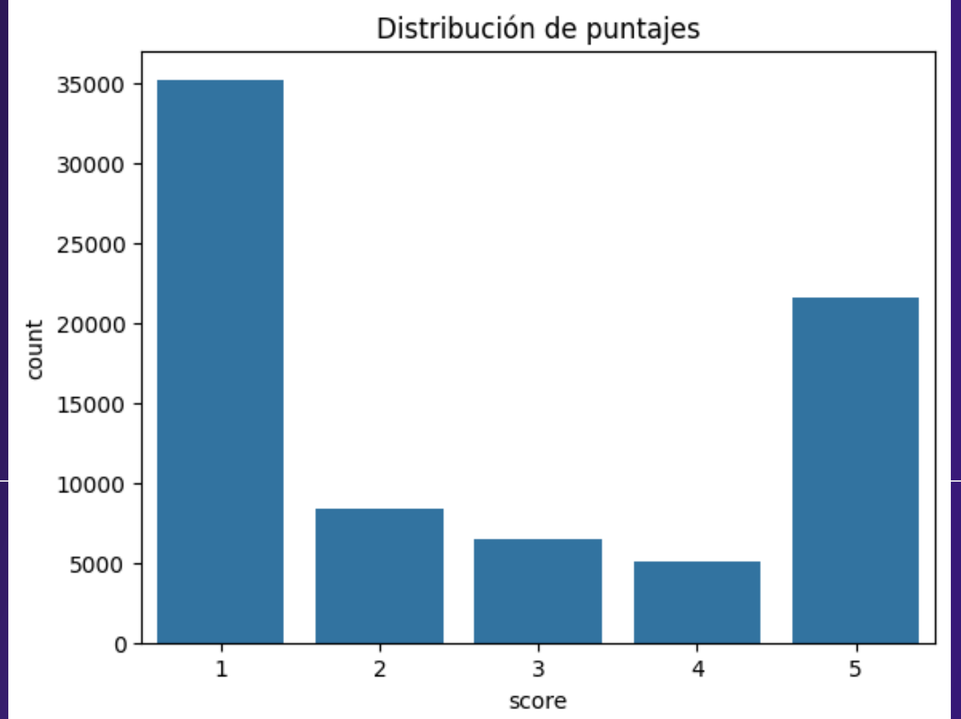
El objetivo de este análisis es aplicar técnicas de Deep Learning para extraer patrones, sentimientos y conocimientos a partir del lenguaje natural presente en los textos. Se buscará entender la satisfacción de los usuarios, detectar puntos de mejora y aportar a la toma de decisiones para la mejora continua de productos digitales.

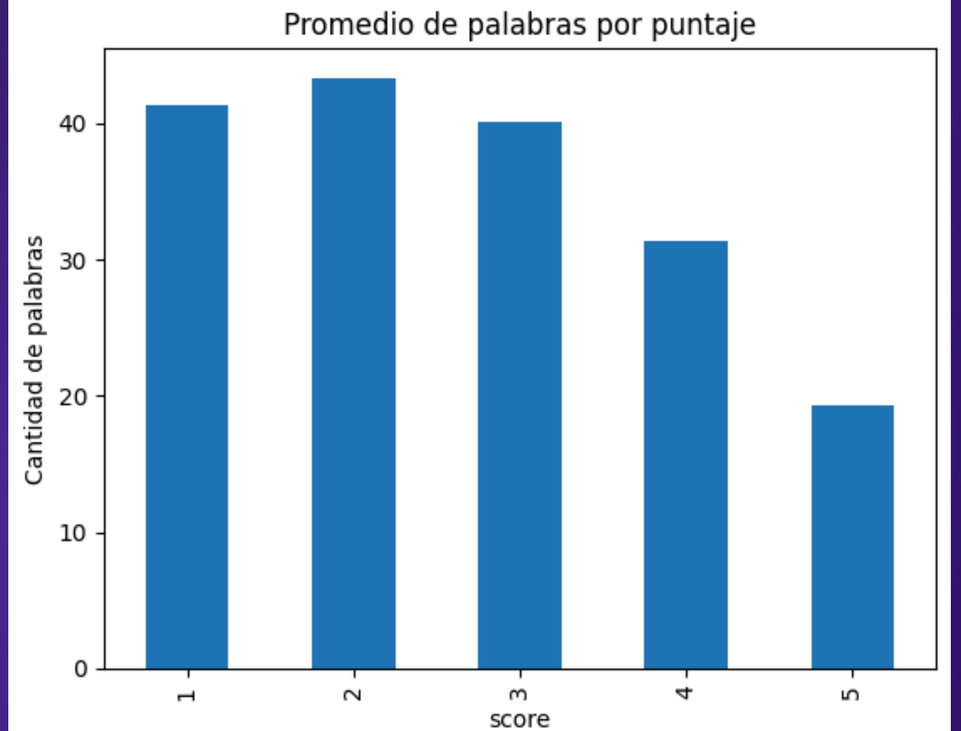
Limpieza de datos y análisis exploratorio

La variable objetivo es score (int. de 1 a 5): indica la calificación que un usuario le dio a la app.

Se observa una distribución desequilibrada de puntajes, la calificación 1 y 5 son las más frecuentes, especialmente las negativas (1)). Esto sugiere que los usuarios tienden a dejar reseñas cuando tienen experiencias muy malas o muy buenas.

Los puntajes bajos contienen más palabras en promedio, los usuarios expresan más detalles al quejarse que al dejar opiniones positivas.





Entrenamiento de red neuronal

```
Epoch 1/5
96/96 ━━━━━━━━━━━━━━━━━━━━ 4s 26ms/step - accuracy: 0.6425 - loss: 0.6418 - val_accuracy: 0.7589 - val_loss: 0.5345
Epoch 2/5
96/96 ━━━━━━━━━━━━━━━━━━━━ 1s 6ms/step - accuracy: 0.7711 - loss: 0.5191 - val_accuracy: 0.8098 - val_loss: 0.4484
Epoch 3/5
96/96 ━━━━━━━━━━━━━━━━━━━━ 1s 4ms/step - accuracy: 0.8376 - loss: 0.4164 - val_accuracy: 0.8767 - val_loss: 0.3428
Epoch 4/5
96/96 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.8834 - loss: 0.3264 - val_accuracy: 0.8982 - val_loss: 0.2963
Epoch 5/5
96/96 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.8930 - loss: 0.2887 - val_accuracy: 0.8970 - val_loss: 0.2827
480/480 ━━━━━━━━━━━━━━━━━━━━ 2s 3ms/step - accuracy: 0.9037 - loss: 0.2723

🔍 Accuracy en test: 0.9043
480/480 ━━━━━━━━━━━━━━━━━━━━ 1s 2ms/step
```

Se eligió una red neuronal simple con 4 capas, eficiente para tareas básicas de clasificación de texto y sin requerir gran capacidad computacional.

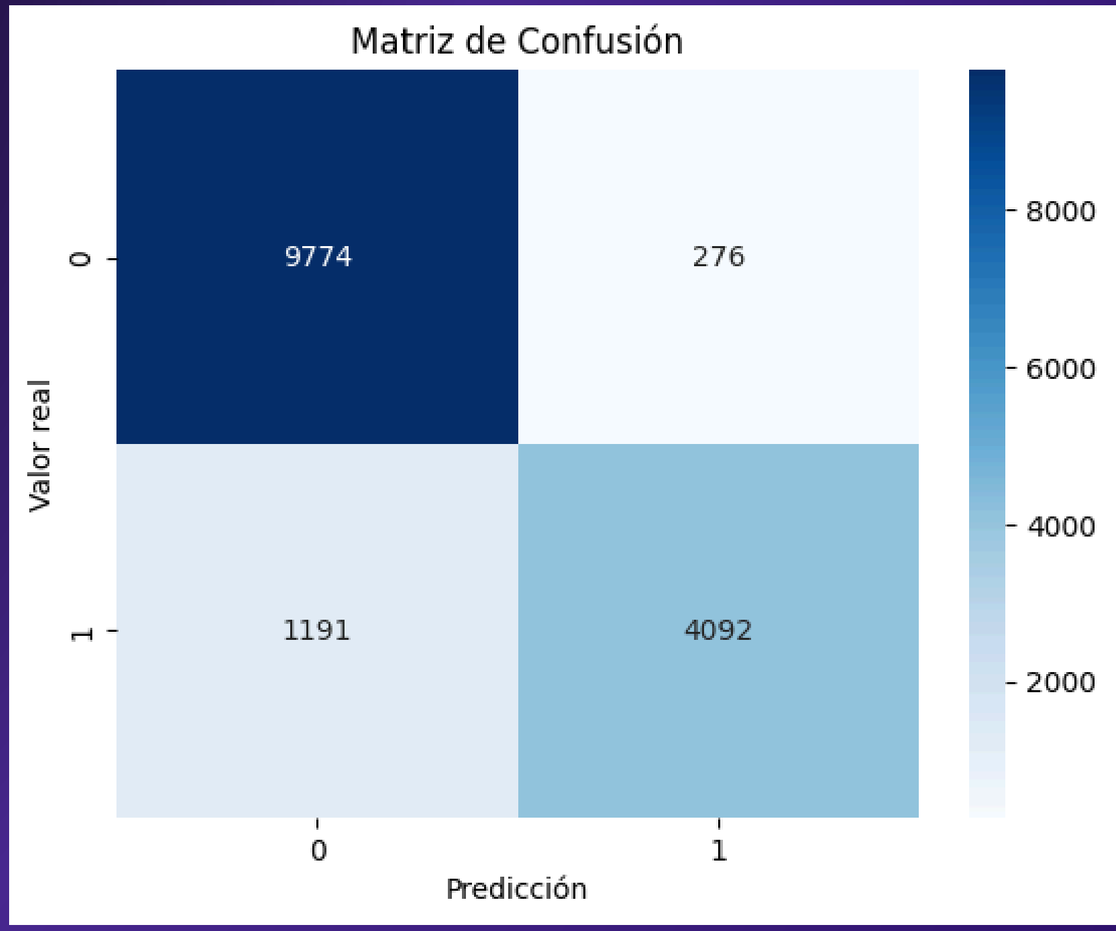
Durante el entrenamiento, se observó una mejora constante en la precisión y la reducción de la pérdida, en el conjunto de entrenamiento y en el de validación. Se alcanzó una precisión final de validación alrededor del 89% y en test del 90%.
Sugiere que no hubo sobreajuste significativo y que el modelo generaliza bien.

Entrenamiento de red neuronal

La matriz de confusión indica:

TN (Verdadero Negativo): 9774
FP (Falso Positivo): 276
FN (Falso Negativo): 1191
TP (Verdadero Positivo): 4092

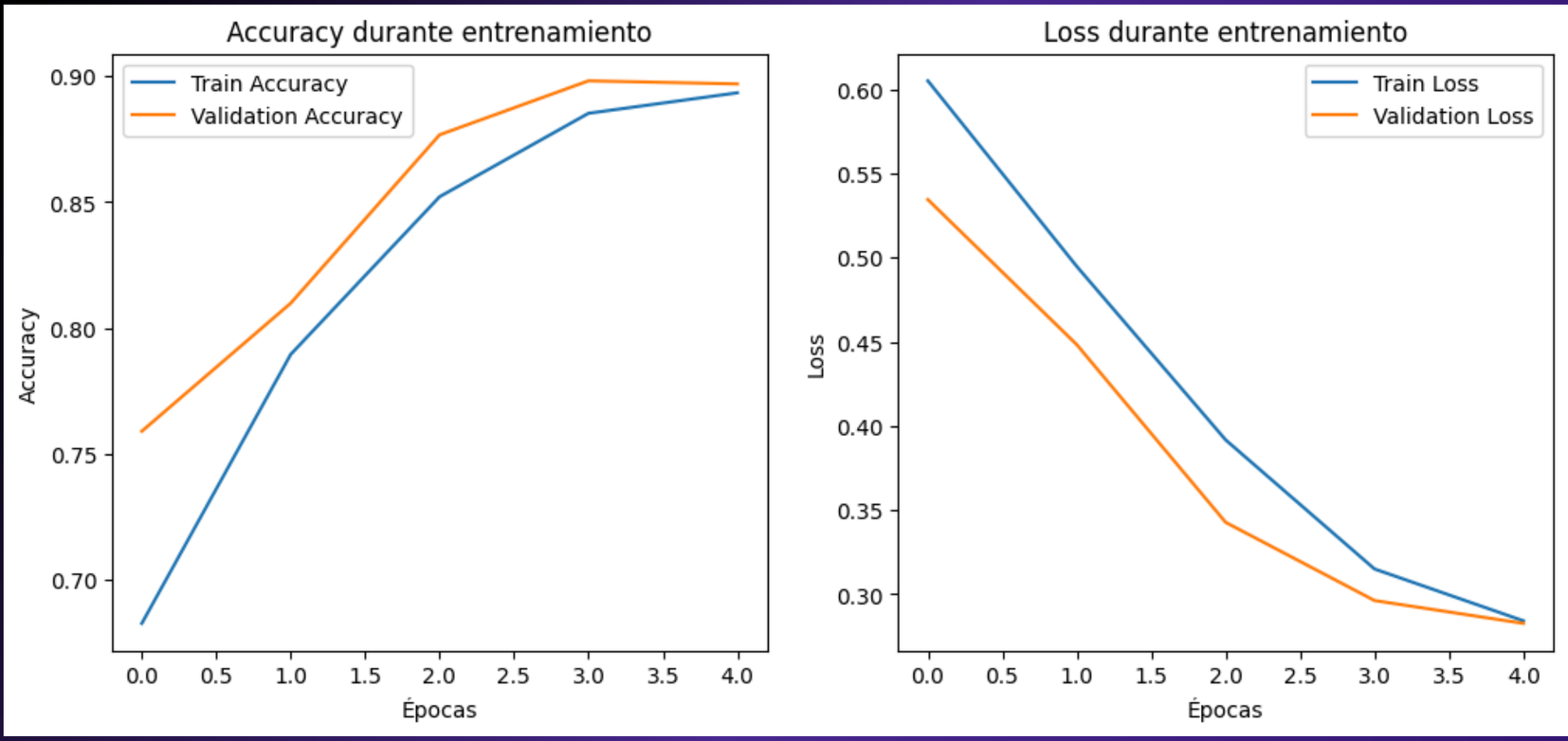
Esto sugiere que el modelo tiene buen desempeño, con un equilibrio razonable entre falsos positivos y falsos negativos, aunque hay una mayor proporción de falsos negativos comparada con falsos positivos.



Comparación

Accuracy:
Train accuracy -> empieza más baja y va aumentando hasta ~ 0.89.
Valid. accuracy → comienza más alta y mejora, llegando a ~ 0.89, luego se estabiliza o incluso baja.

El modelo aprende bien y generaliza bien durante las primeras épocas, pero después podría estar sobreajustando, ya que la valid. accuracy no mejora y train accurancy sigue subiendo.



Loss:
Train loss → disminuye de 0.60 a 0.30, indica que el modelo está aprendiendo y ajustando sus pesos.
Valid. loss → baja rápidamente, a ~ 0.30, luego se mantiene estable y no disminuye mucho después de la época 2.

Esto respalda la idea de que el modelo aprende bien, pero a partir de la época 2-3 se estabiliza. La diferencia que se ve entre train loss y valid. loss no es muy grande, lo que indica que no hay un sobreajuste fuerte.

Conclusiones sobre el dataset elegido

Se entrenó una red neuronal simple de 4 capas para clasificar reseñas de Amazon como positivas o negativas, logrando una precisión aceptable. El modelo logró buen desempeño general, un accuracy del 90%, un buen resultado para un problema de clasificación binaria.

Algunas instancias positivas no fueron detectadas (falsos negativos) y mejorar en este aspecto podría ser importante dependiendo del contexto y la criticidad del error.
Como se mencionó anteriormente, el modelo mejora y luego se estabiliza, señal de un posible inicio de sobreajuste leve.

Este estudio resulta útil para automatizar el análisis de opiniones de clientes, ayudando a empresas a detectar rápidamente niveles de satisfacción, identificar productos con problemas recurrentes y mejorar la experiencia del usuario.

GRACIAS

CODERHOUSE
Comisión 61790
Alumna: Méndez, Julieta Milagros

2025